

Wine Quality Classification

CSC44700 - Final Project

Alice Liu









Build a model to predict the quality of wine based on various chemical attributes of the wine





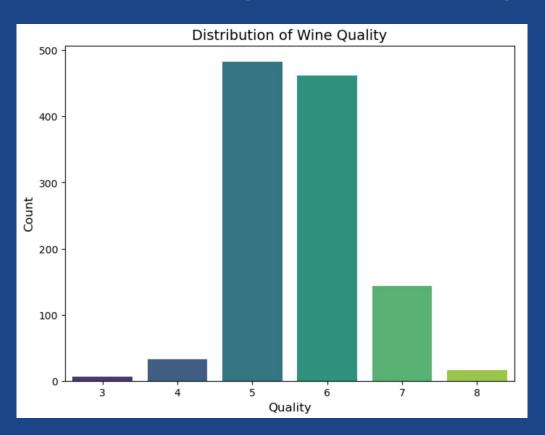




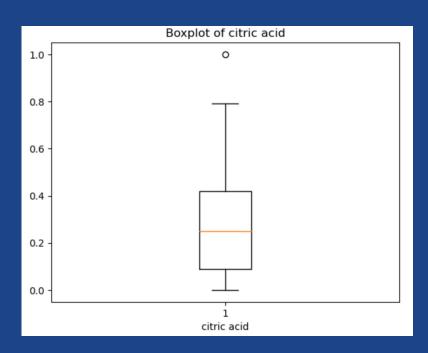
- Shape: (1143, 13)
- Columns: 13
 - 'fixed acidity', 'volatile acidity', 'citric acidity', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol', 'Id'
 - Target: 'quality'
- No NULL

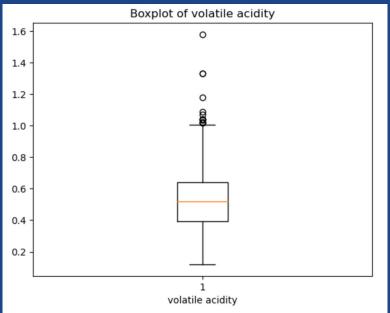


Dataset Distribution of Unique Value from 'quality'



🕨 Initial EDA and Cleaning 🖪

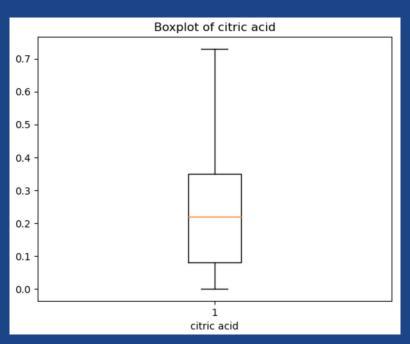


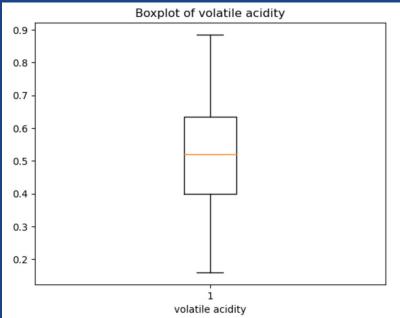


Initial EDA and Cleaning

```
# DROP Outliers for all columns using IQR
def drop outliers(df):
    df new = pd.DataFrame()
    for col in df.columns:
        q1 = df[col].quantile(0.25)
        q3 = df[col].quantile(0.75)
        IOR = a3 - a1
        lower_bound = q1 - 1 * IQR # 1.5 and 2 didn't work; 1 worked the best for threshold
        upper_bound = q3 + 1 * IQR # 1.5 and 2 didn't work; 1 worked the best for threshold
        # drop outliers
        df \ new[col] = df[(df[col] >= lower bound) & (df[col] <= upper bound)][col]
    return df_new
df_clean = drop_outliers(df)
df_clean = df_clean.dropna()
plt_boxplots(df_clean)
```

Initial EDA and Cleaning











• Shape: (633, 12)

• Columns: 12

'fixed acidity', 'volatile acidity', 'citric acidity', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol'

o Target: 'quality'



ŀ

1

0.048

-0.65

-0.048

0.00047

-0.65

0.024

-0.036

0.074

fixed acidity -

volatile acidity

residual sugar

free sulfur dioxide

total sulfur dioxide

citric acid -

chlorides

density

sulphates

alcohol

quality -

-0.096

-0.65

0.046

Heatmap of Feature Correlation

-0.036

-0.048

1

0.64

-0.039

0.059

-0.16

-0.096

0.00047

0.64

-0.074

-0.28

More EDA

0.046

0.048

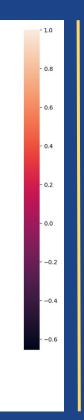
-0.65

-0.098

1

표

-0.098





- Alcohol vs Quality
- Sulphates vs Quality
- Citric Acid vs Fixed Acidity
- Density vs Fixed Acidity
- Free SO2 vs Total SO2

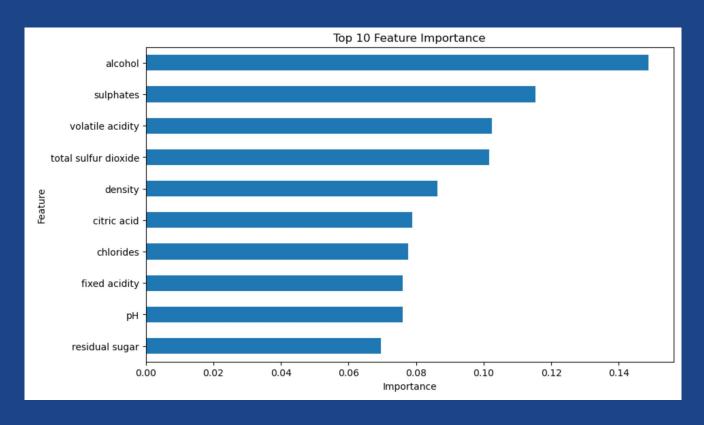
Correlation

- Fixed Acidity vs pH
- Citric Acid vs Volatile Acidity
- Alcohol vs Density

ŀ

FEATURE IMPORTANCE





MODELING









Logistic Regression

Decision Trees

Random Forest

max_iter = 1000. random_state = 42

criterion = gini,

n_estimators = 200. random_state = 42









SGD Classifier

Gradient Boosting

KNN

Extra Trees

n_neighbors = 5 n_neighbors = 10 n_estimators = 200

n_estimators = 200. learning_rate = 0.1, random_state = 42

- loss = 'hinge', random_state = 42
- loss = 'modified_huber', random_state = 42

► MODELING - Feature Selection •

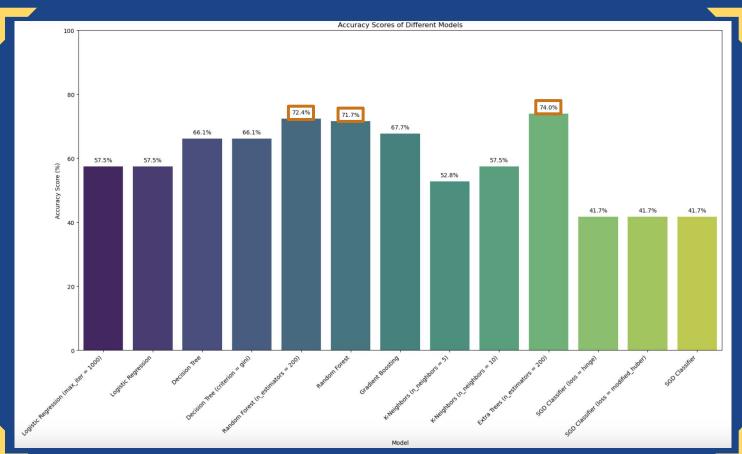
```
In [67]: # using TOP 3 FEATURES FROM FEATURE IMPORTANCE <- 72% in RF, 74% in ExtraTrees
X_clean = df_clean[['alcohol', 'sulphates', 'volatile acidity']]
y_clean = df_clean['quality']

X_train_clean, X_test_clean, y_train_clean, y_test_clean = train_test_split(X_clean, y_clean, test_size = 0.2, random_state = 42)</pre>
```



EVALUATION





HIGHEST: EXTRA TREES CLASSIFIER



POINTS OF IMPROVEMENTS



Feature Engineering

Explore additional feature engineering techniques (e.g., scaling, more effective handling of outliers, etc.)



Hyperparameter Tuning

Adjust the parameters, experiment using GridSearchCV



Model Selection

Try different models to see if they offer better performance, consider advanced ensemble techniques (e.g., stacking, blending multiple models)





THANK YOU!

Project GitHub Link: https://github.com/AliceLiu17/Wine-Quality-Prediction